基于结构引导的多模态虚假信息检测方法

谢丽霞[1,2]，罗丹[1]，杨宏宇[1,2*]，朱亚萍[2]
（1.中国民航大学计算机与人工智能学院，天津 300300；2.中国民航大学安全科学与工程学院，天津 300300）

摘要：针对现有多模态虚假信息检测方法在跨模态语义交互过程中缺乏传播结构约束，以及固定融合策略难以适应样本级模态质量差异的问题，提出一种基于结构引导与样本级自适应融合的虚假信息检测方法（Structure-Guided Multimodal Adaptive Fusion，SGMAF）。首先，从信息传播图中提取节点影响力、传播时序、边类型及节点度等拓扑特征，构建结构偏置矩阵，并将其作为加性偏置注入图文跨模态注意力计算过程，对跨模态注意力分布进行约束与重加权。其次，提出一种样本级自适应多模态融合方法，根据文本、图像及传播结构在当前样本中的信息质量动态生成模态融合权重，以降低低质量模态对模型判别结果的干扰。最后，在Pheme与微博两个公开数据集上进行了实验验证。实验结果表明，所提方法在Accuracy、Precision、Recall及F1-score等指标上均优于现有主流基线模型，其中在Pheme与微博数据集上的准确率分别达到90.91%和96.27%，验证了结构引导机制与动态融合策略的有效性。

关键词：虚假信息检测；多模态学习；结构引导机制；自适应融合；信息传播网络
中图分类号：TP393.08
文献标志码：A
doi：10.11959/j.issn.1000

Structure-Guided Multimodal Misinformation Detection Method

Xie Lixia[1,2], Luo Dan[1], Yang Hongyu[1,2*], Zhu Yaping[2]
1. School of Computer Science and Artificial Intelligence, Civil Aviation University of China, Tianjin 300300, China
2. School of Safety Science and Engineering, Civil Aviation University of China, Tianjin 300300, China

Abstract: Current multimodal misinformation detection methods lack propagation structure constraints on cross-modal semantic interaction, and fixed fusion strategies fail to accommodate sample-level variations in modal quality. To address these issues, a Structure-Guided Multimodal Adaptive Fusion method (SGMAF) was proposed. First, topological features including node influence, propagation timing, edge types, and node degrees were extracted from the information propagation graph to construct a structural bias matrix, which was injected into the text-image cross-modal attention computation to constrain and reweight the attention distribution. Second, a sample-level adaptive multimodal fusion method was designed to dynamically generate modality fusion weights based on the information quality of text, images, and propagation structures within each sample, thereby mitigating the interference of low-quality modalities. Experimental evaluations were conducted on two public datasets, Pheme and Weibo. The results showed that SGMAF outperformed mainstream baseline models across Accuracy, Precision, Recall, and F1-score, achieving 90.91% and 96.27% accuracy on Pheme and Weibo, respectively. These results validate the effectiveness of the structure-guided mechanism and the dynamic fusion strategy.

收稿日期：2026–05–22；**修回日期：**2026–07–07
通信作者：杨宏宇，yang_hy2006@126.com
基金项目：国家自然科学基金民航联合研究基金重点项目(U2433205)；国家重点研发计划项目(2025YFB3110500)
Foundation Items: Civil Aviation Joint Research Fund Project of the National Natural Science Foundation of China (No. U2433205); National Key Research and Development Program of China (2025YFB3110500)

Key words: misinformation detection, multimodal learning, structure-guided mechanism, adaptive fusion, information propagation network

0 引言

随着在线社交平台的快速发展，信息传播模式已由单一文本逐步演变为融合文本内容、视觉图像与用户交互行为的多模态传播范式[1]。在真实的社交媒体场景里，一条新闻不仅有文本和配图这些内容，还会伴随用户转发、评论还有引用这类传播行为，这些信息共同影响新闻真假的判断。准确检测虚假信息，不仅能阻止虚假信息在社交网络里快速传开，提高网络空间的信息安全程度，同时也是后面做舆情治理、传播溯源还有社会风险分析的重要基础[2]。随着生成式人工智能和对抗规避技术的发展，虚假信息的组织与传播策略日益复杂[3]。造谣者通过文本伪装、图像篡改以及“真实图像加虚假文本”等方式增强迷惑性，使得基于浅层语义匹配的检测方法难以识别深层矛盾[4]；同时，虚假信息往往伴随异常扩散模式，不同层级的传播行为也隐含着重要的真实性线索。此外，图像噪声、文本噪声和传播树稀疏等问题进一步加剧了检测复杂性[5]。这些因素使传统依赖单一模态特征或固定融合策略的方法难以满足复杂社交环境下的鲁棒检测需求。

近年来，视觉-语言预训练模型（Vision-Language Pretrained Models, VLPMs）通过大规模图文对齐学习，显著提升了跨模态语义关联的建模能力，推动了图像描述、视觉问答等多模态任务的发展[6-7]。此外，图注意力网络（Graph Attention Network, GAT）等方法通过对用户信息，例如转发、评论等交互行为进行传播图建模，实现了传播拓扑结构与节点关联特征的联合学习，从而能够有效捕获虚假信息扩散过程中的异常传播模式[8]。VLPMs与GAT等方法的深入应用，推动了虚假信息检测从单模态浅层语义分析向多模态深层语义与传播结构协同建模的演进[9]。尽管现有多模态方法在图文语义建模方面取得了一定进展，但仍存在两个关键局限。其一，大多数方法将跨模态语义交互与传播结构建模分离，仅在后期融合阶段简单结合结构特征与图文语义，导致跨模态对齐缺乏传播拓扑的显式约束，从而容易将图文表面匹配但传播行为异常的虚假信息误判为真实信息[10]。其二，现有融合机制普遍采用基于全局统计的固定权重策略，难以适应不同样本之间显著的模态质量差异[11]。当图像噪声、文本噪声增强或传播结构不完整时，低质量模态容易在融合过程中被过度放大，导致模型鲁棒性明显下降。

针对上述问题，本文提出了一种基于结构引导与样本级自适应融合的社交媒体虚假信息检测方法（Structure-Guided Multimodal Adaptive Fusion, SGMAF）。首先，为增强传播结构对跨模态语义交互的约束能力，本文设计结构引导的跨模态融合注意力机制，从传播拓扑图中提取节点影响力、时序关系、边类型及节点度等多维拓扑特征，构建结构偏置矩阵，并将其作为加性偏置显式注入图文跨模态注意力计算过程，使传播行为能够为图文交互提供结构先验，从而减少模型对表层语义一致性伪装的误判。其次，针对不同样本模态质量波动较大的问题，本文提出一种样本级自适应多模态融合方法，通过动态感知文本、图像与传播图的信息质量，自适应生成模态融合权重，以抑制低质量噪声模态的干扰并强化高判别力模态的贡献。最终，本文模型通过联合建模传播拓扑结构与多模态语义特征，实现了对社交媒体虚假信息的有效识别。

（1）构建了一种基于传播结构引导的跨模态注意力交互机制。该机制将传播拓扑中的节点影响力、传播时序及交互关系等结构信息转化为显式偏置，引导文本与图像之间的跨模态注意力计算，实现传播行为对图文交互过程的约束与校正。

（2）提出一种样本级自适应多模态融合方法。该方法能够根据当前样本各模态的信息质量动态调整融合权重，有效缓解固定融合策略在模态缺失、噪声干扰等复杂场景下的鲁棒性不足问题。

（3）提出SGMAF虚假信息检测框架，并在Pheme与微博两个真实社交媒体数据集上进行了广泛实验。结果表明，所提方法在准确率、精确率、召回率及F1分数等评价指标上均优于现有主流基线模型，验证了结构引导机制与自适应多模态融合方法的有效性。

1　相关工作

1.1　基于内容的虚假信息检测

早期虚假信息检测主要视作文本分类任务，通常依赖人工构建的词频统计、情感词典及语言风格特征，并结合逻辑回归或随机森林等传统机器学习分类器进行判别[12]。此类方法高度依赖专家先验知识，且离散的统计特征丢失了语句的上下文顺序，导致浅层特征难以刻画复杂的语义逻辑关系。为了克服特征工程所面临的挑战，表示学习的发展推动了研究范式向深度神经网络的迁移。卷积神经网络（Convolutional Neural Network, CNN）与循环神经网络（Recurrent Neural Network, RNN）及其衍生模型被广泛用于文本特征的自动抽取，分别负责提取其中的局部短语模式与长距离时序依赖[13]。然而，CNN 的局部注意难以跨越长文本建立全局关联，而 RNN 在处理长链路文本时存在梯度衰减与信息遗忘问题，这使得它们检测高度仿真的虚假信息时性能不足。此后，多尺度网络结构与注意力机制被逐步引入，用以增强模型对文本内容特征的建模能力[14]。预训练语言模型应用于文本表征范式，通过海量语料库的深层上下文学习，BERT 与 RoBERTa 等架构在现有评估中提升了纯文本检测的基线性能[15]。网络虚假信息正演化出极具隐蔽性的“语义伪装性”。面对被刻意构造以混淆视听的文本表层逻辑，过度依赖单一文本映射的模型或许极易陷入决策误区[16]。在线信息流具有高度碎片化属性，现阶段内容交互往往深度绑定图像或视频等多模态载体。若在分析中强制剥离视觉维度，这不仅可能割裂完整语境，更会阻碍系统捕获“图文不匹配”这一核心线索[17]。单模态感知盲区已难以规避[18]。上述局限正促使研究范式转向多模态协同建模。

1.2　基于传播结构的虚假信息检测

除内容语义外，信息在社交网络中的传播拓扑与扩散时序同样携带独立的判别价值。研究发现，虚假信息在扩散初期常表现出异常的爆发速率与特定的层级深度[19]。早期检测手段多依赖人工定义的统计特征，如传播树深度、最大宽度及节点度分布等离散指标[20]。此类方法的瓶颈在于动态级联传播被简化为静态宏观指标，细粒度交互路径与时序信息随之不可逆丢失；而显式特征在对抗性行为面前的脆弱性也暴露无遗，自动化手段可轻易构造“结构伪装”，使规则直接失效。图神经网络（Graph Neural Network, GNN）在这一困境下被引入，并逐步确立了在高阶拓扑分析中的主导地位[21]。其端到端的消息传递机制绕开了人工特征设计，在聚合局部邻居与全局拓扑时对结构噪声展现出更强的抵抗力。代表性工作 EBGCN 通过贝叶斯图卷积网络刻画交互边的不确定性，相关文献证实了其捕获潜在虚假模式的有效性[22]。问题在于纯结构建模将内容语义完全剥离后，模型便丧失了对传播底层驱动逻辑的解析能力，传播行为的演化归根结底受语义与受众心理的深度控制。当遭遇高信誉源无意转发或水军定向扩散时，缺乏语义锚定的图结构所输出的表征易产生误导，进而诱发大面积检测失效。将传播网络的高阶拓扑与节点多模态语义深度耦合，在这一背景下已不仅是技术路线偏好，也是突破当前性能瓶颈的先决条件。

1.3　多模态虚假信息检测

多模态特征的整合方式决定了当前检测框架的能力边界[23]。早期跨模态对齐范式所依赖的前提假设正被重新审视。以潜在空间对齐为路径的方法，如 MVAE[24] 和 SAFE[25]，分别通过变分自编码器联合推导视觉与文本的共享表征，以及在统一子空间内显式量化图文语义相似度。这类方法的共同之处在于隐含一种同源性预设，认为文本与图像在信息传递上保持严格一致。然而在真实传播环境中，真实图像与伪造文本的拼接所引发的跨模态逻辑断裂，恰是虚假信息的重要表征。此时，对语义绝对一致性的前置依赖反而可能使模型在关键场景下丧失判别力。更深层的问题在于，异质模态的强制对齐将来源相异、表征逻辑不同的特征流生硬映射至单一共享空间，极易造成模态特有信息的流失，并在深层削弱整体语义表达。这一认识促使研究者开始寻找能够在保持模态差异的前提下实现有效交互的新路径。为克服上述局限，以 CLIP（Contrastive Language‑Image Pre-training）为代表的视觉-语言预训练模型通过大规模对比学习实现跨模态语义深度对齐，已被用于提取多模态虚假信息的统一特征表示[26]。然而，CLIP 的通用对齐特征更侧重于宏观语义匹配，在捕捉特定新闻场景下细粒度的“图文逻辑矛盾”时仍显不足，且对旨在误导语义对齐的对抗性样本时鲁棒性受限。此类方法虽能量化图文一致性，但在跨模态交互的整个计

算图内，传播拓扑信息通常处于缺失状态，导致模型忽视信息演化中累积的结构影响力。

近年来，相关研究开始将图结构引入多模态框架中，MFAN[27]通过多头注意力联合编码文本、图像与结构特征，但其内部呈“松散耦合”：图卷积模块局限于邻居特征的独立聚合，跨模态注意力计算则在单节点内部孤立执行，传播拓扑对多模态交互的显式引导路径被切断。这使得模型在评估不同传播层级节点的差异化校验权重感知不足。语义空间统一与策略优化构成另一分支。NLIN将异质特征映射至自然语言空间并依托推理框架判别[28]，局部缓解了模态异构性问题，但推理仍由内容语义主导。面对“高信誉源无意转发虚假信息”等场景，缺乏结构层面的显式纠偏，系统易受源权威性误导。CLFFRD引入课程学习以强化局部语义映射[29]，然而拓扑维度信息依旧边缘化；逐步递进的训练范式虽巩固了收敛稳定性，却未触及鲁棒性衰减的根因，样本间剧烈的模态质量波动。当前多模态虚假信息检测进一步向混合专家建模与大规模多模态模型驱动的推理范式演进。以MIMoE-FND为代表的方法[30]，通过模态交互的混合专家机制显式建模不同模态之间的专长分工。同时，基于大规模多模态模型的方法也在不断发展，主要包括参数冻结、参数微调和推理增强三类范式，其中，Human-Like[31]框架采用Qwen-VL-Max视觉-语言大模型完成图文联合推理，并利用Qwen-Turbo纯文本大模型生成反思式摘要，通过大模型与小模型协同机制增强多模态内容理解与推理能力。这类方法主要依赖图文内容的语义分析与跨模态推理实现虚假信息检测，对传播拓扑结构等行为信息的显式建模相对有限，同时在实际部署中通常伴随较高的推理成本和更复杂的系统流程。针对上述不足，本文从传播结构约束与样本级自适应融合两个层面切入，提出结构引导的样本级自适应多模态融合模型，赋予传播拓扑主动权，利用图结构先验干预图文注意力分配，并依据模态质量推导自适应门控权重，以提升模型在复杂社交场景中的检测能力。

2　方法设计

2.1　多模态虚假信息问题形式化定义

在社交媒体环境中，一条信息通常同时包含文本内容、视觉信息以及用户传播行为，这些信息共同反映了新闻的真实性。因此，可以将虚假信息检测问题设定为一个多模态分类任务。设社交媒体数据集为$D=\{(X_i,y_i)\}_{i=1}^{N}$，其中$X_i$表示第$i$条信息样本，$N$为数据集样本总数，$y_i\in\{0,1\}$表示样本标签，当$y_i=1$时表示虚假信息，当$y_i=0$时表示真实信息。在实际场景中，每条新闻样本通常包含三类信息：新闻文本、配图以及用户传播关系。因此可以将输入样本表示为：

$$X_i=(T_i,I_i,G_i) \tag{1}$$

其中，T_i表示信息的文本序列；I_i表示信息对应的图像；G_i表示由用户转发行为构成的传播图结构，其中V_i为用户节点集合，E_i为转发关系集合。本文的目标是学习一个映射函数：

$$F:(T_i,I_i,G_i)\rightarrow y_i \tag{2}$$

该函数能够综合利用文本语义、视觉内容以及传播结构三类信息，输出信息为虚假信息的概率。为实现上述目标，本文提出SGMAF模型，通过多模态特征编码、结构引导跨模态交互及样本级自适应多模态融合方法三个维度的协同，提升模型在多源异构信息中完成特征整合与稳健判别的能力。

2.2　SGMAF总体框架

本文提出的SGMAF模型整体框架如图1所示。模型的核心思想是利用传播拓扑结构作为先验知识，显式引导文本与图像之间的跨模态语义交互，同时引入样本级自适应多模态融合方法，通过自适应调节机制动态评估各模态的信息质量并据此调整融合权重。SGMAF模型由以下四个核心部分组成：

（1）多模态特征编码：对文本、图像、传播图分别编码，得到统一维度的表征；

（2）结构引导跨模态交互：从传播图中提取拓扑特征构造结构偏置矩阵，将其注入图文跨模态注意力，获得结构增强的表征；

（3）样本级自适应多模态融合：根据样本质量动态生成各模态权重，加权融合得到最终表征；

（4）分类决策：根据最终表征，输出预测概率。

上述模块中，除分类决策为固定组件外，前三个部分的核心设计在消融实验中分别通过移除结构偏置（-w/o B）、移除自适应多模态融合（-w/o A）、移除图结构模态（-w/o G）、移除视觉模态（-w/o V）等变体进行有效性验证。模型具体工作流程如下四个阶段：

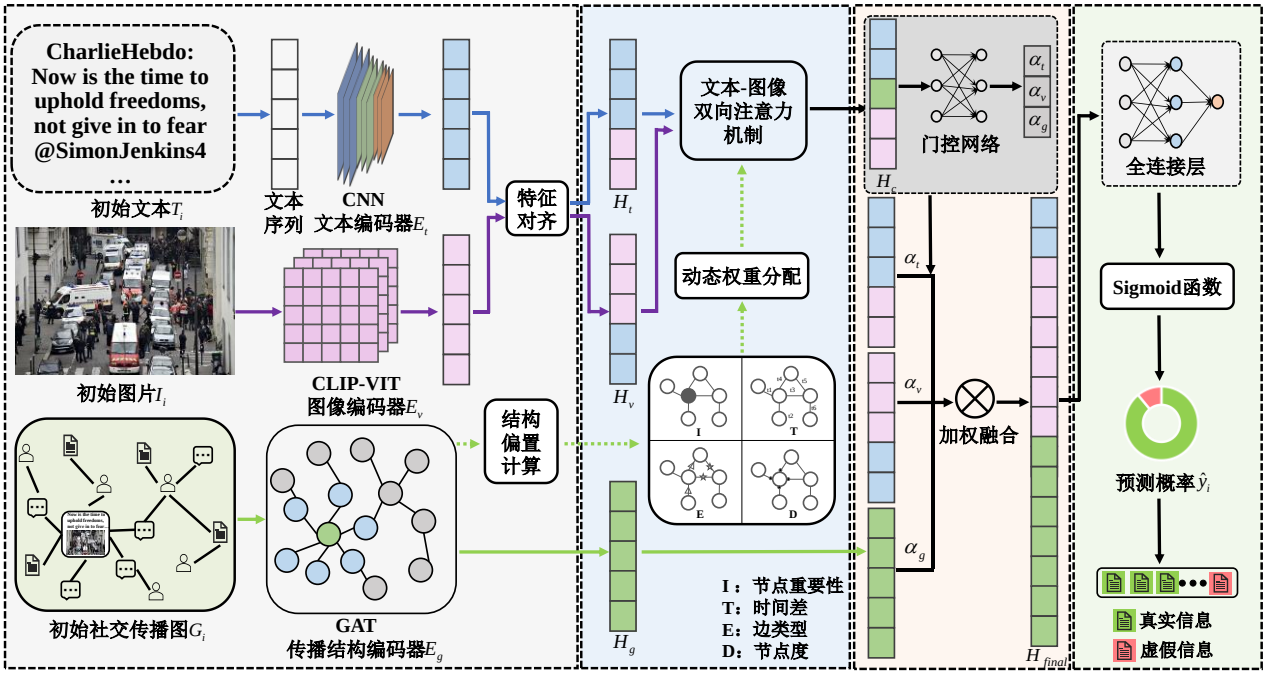


图1　基于结构引导的多模态虚假信息检测模型框架

（1）在多模态特征编码阶段，输入原始三元组(T_i,I_i,G_i)，分别采用文本编码器E_t、图像编码器E_v和传播结构编码器E_g对原始输入进行特征抽取，将三类异构数据映射至统一的向量语义空间，得到文本表示$\boldsymbol{H}_t \in \mathbb{R}^d$、视觉表示$\boldsymbol{H}_v \in \mathbb{R}^d$和结构表示$\boldsymbol{H}_g \in \mathbb{R}^d$；

（2）从传播图G_i中提取多维拓扑特征构造结构偏置矩阵$\boldsymbol{B}$，将其作为加性项注入文本与图像之间的跨模态注意力计算过程。通过传播结构的显式约束，对图文交互中的注意力分布进行重加权，获得结构增强的跨模态联合表示$\boldsymbol{H}_c \in \mathbb{R}^d$；

（3）将跨模态联合表示$\boldsymbol{H}_c$与结构表示$\boldsymbol{H}_g$拼接后输入轻量级门控网络，动态生成模态权重系数α_t，α_v，α_g。该权重向量能够感知当前样本各模态的信息质量，并指导文本基础表示$\boldsymbol{H}_t$、图像基础表示$\boldsymbol{H}_v$及结构基础表示$\boldsymbol{H}_g$的加权融合，生成最终的多模态表征$\boldsymbol{H}_{\text{final}}$；

（4）将融合表征$\boldsymbol{H}_{\text{final}}$送入由单层全连接网络和Sigmoid激活函数构成的分类器，输出预测概率$\hat{y}_i$，完成对输入虚假信息的二分类判别。

2.3　多模态特征编码模块

异构数据的同源映射构成了初始处理阶段，多模态特征编码模块尝试在统一语义空间内执行原始输入的向量化重构，提供高质量的特征基础，以进行后续的跨模态交互。

2.3.1　文本特征编码

虚假信息的文本通常兼具整体语义倾向与局部诱导特征。例如，虚假信息文本常在全局层面模仿权威语体以降低读者警惕性，同时在局部穿插“震惊!”、“速看!”等情绪化短语以刺激传播。为同时捕获上述两类特征，本文设计了一种全局-局部双分支文本编码架构。

给定输入词序列$T_i=\{w_1,w_2,\cdots,w_n\}$，首先通过词嵌入层获得词向量序列$\boldsymbol{X} \in \mathbb{R}^{n \times d_e}$，并叠加正弦位置编码以注入序列顺序信息。全局语义分支采用$L_t=6$层Transformer编码器对序列进行上下文建模，取首token[CLS]对应的输出向量$\boldsymbol{h}_{\text{cls}} \in \mathbb{R}^{d_e}$作为全局语义表示，该向量经自注意力机制聚合了整段文本的上下文信息。局部特征分支并行采用三组一维卷积神经网络，卷积核尺寸分别设为$k_1=2$、$k_2=3$、$k_3=4$，以捕捉不同粒度的短语模式，每组卷积的输出通道数均设为$c=128$。对每个卷积核的输出序列分别施加时序最大池化，获得定长向量$\boldsymbol{h}_{\text{cnn}}^{(1)},\boldsymbol{h}_{\text{cnn}}^{(2)},\boldsymbol{h}_{\text{cnn}}^{(3)} \in \mathbb{R}^{128}$，将其拼接后得到局部特征表示$\boldsymbol{h}_{\text{cnn}} \in \mathbb{R}^{384}$。最终，将全局语义向量与多尺度局部向量拼接，并通过一个线性投影层映射至统一维度$d=256$，得到文本模态的总体表示:

$$\boldsymbol{H}_t=\boldsymbol{W}_t \cdot [\boldsymbol{h}_{\text{cls}} \| \boldsymbol{h}_{\text{cnn}}] + \boldsymbol{b}_t \tag{3}$$

其中 $\boldsymbol{W}_t \in \mathbb{R}^{256\times(d_e+384)}$ 为可学习权重矩阵，$\boldsymbol{b}_t$ 为偏置项，$\|$ 表示向量拼接。该设计使文本编码器既能理解新闻的整体语义倾向，又能敏锐捕捉虚假信息中具有强判别力的局部诱导性短语。

2.3.2 图像特征编码

视觉内容是虚假信息的重要载体。造谣者常通过篡改图像、盗用无关图片或添加误导性标注等手段增强信息的欺骗性。为获取与文本语义空间天然对齐的视觉表征，本文采用基于对比学习预训练的CLIP视觉编码器提取图像特征。选择CLIP-ViT-B/32作为图像编码器 E_v。CLIP通过最大化文本和图像对的正样本相似度、最小化负样本相似度的大规模对比学习，使其视觉编码器习得的特征，然后与自然语言语义空间高度对齐。给定输入图像 V_i，首先将其缩放到224 × 224分辨率并切分为固定大小的图像块，经ViT编码器处理后，取[CLS]标记对应的输出向量 $\boldsymbol{v}_{\text{cls}} \in \mathbb{R}^{768}$ 作为图像的全局语义表征。随后，通过线性投影层将其映射至与文本特征相同的向量空间:

$$\boldsymbol{H}_v = \boldsymbol{W}_v \cdot \boldsymbol{v}_{\text{cls}} + \boldsymbol{b}_v \tag{4}$$

其中 $\boldsymbol{W}_v \in \mathbb{R}^{256\times 768}$ 为投影矩阵，$\boldsymbol{b}_v$ 为偏置项。CLIP特征与生俱来的跨模态对齐属性，为后续文本与图像的深度语义交互提供了良好的初始化基础。

2.3.3 图结构特征编码

社交网络中，虚假信息的扩散轨迹往往呈现特定的结构特征：传播链路更短、爆发速率更剧烈、且向关键节点高度聚集。这些异常拓扑模式构成了后续特征编码的出发点。本文采用图注意力网络对传播图进行编码，其选择依据在于不同用户节点的影响力具有差异性，中心节点常引发拓扑放大效应，而边缘节点或自动化水军的结构贡献则相对有限。传统图卷积对邻居节点的无差别聚合，难以捕捉这种影响力落差。图注意力网络通过自适应推导局部权重分配，将聚合重心偏向高影响力节点，从而更精确地刻画异常传播的拓扑范式。

传播图的拓扑结构以单条新闻为中心节点，与之产生转发或评论等交互行为的用户作为邻居节点，共同构成信息传播子图。本文采用 $L_g = 2$ 层GAT进行消息传递，每层配置 $H = 4$ 个注意力头，每头隐层维度设为 $d_h = 64$，拼接后每层输出维度为256。第 l 层节点 i 的特征更新公式为：

$$\boldsymbol{h}_i^{(l+1)} = \Big\|_{k=1}^{H} \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^k \boldsymbol{W}^k \boldsymbol{h}_j^{(l)}\right) \tag{5}$$

其中 $\|$ 表示多头输出的拼接操作，$\mathcal{N}(i)$ 为节点 i 的邻居集合，W^k 为第 k 个注意力头的可学习权重矩阵，σ 为ELU激活函数。注意力权重 α_{ij}^k 通过以下公式计算：

$$\alpha_{ij}^k = \frac{\exp\left(\text{LeakyReLU}\left(a^T[\boldsymbol{W}^k\boldsymbol{h}_i \| \boldsymbol{W}^k\boldsymbol{h}_j]\right)\right)}{\sum_{u\in\mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(a^T[\boldsymbol{W}^k\boldsymbol{h}_i \| \boldsymbol{W}^k\boldsymbol{h}_u]\right)\right)} \tag{6}$$

其中 a 为注意力参数向量。经两层GAT传播后，采用注意力池化而非简单平均池化对全部节点表示进行聚合，以赋予关键节点更高的话语权：

$$\boldsymbol{H}_g = \sum_{i\in\mathcal{U}_i} \beta_i \cdot \boldsymbol{h}_i^{(L)} \tag{7}$$

其中池化权重 β_i 由一个小型评分网络生成：$\beta_i = \text{softmax}(\boldsymbol{u}^T \cdot \tanh(\boldsymbol{W}_p \boldsymbol{h}_i^{(L)}))$，$u$ 和 W_p 均为可学习参数。最终得到的 $\boldsymbol{H}_g \in \mathbb{R}^{256}$ 即为整条新闻的传播结构表示，它凝练了传播拓扑中的关键结构模式。

2.4 结构引导跨模态交互机制

传统的跨模态注意力机制在计算文本与图像的内容关联时，通常仅依赖模态内容特征，忽视了传播行为这一重要先验信息。然而，在社交媒体场景中，传播结构往往隐含着信息真实性的判别线索，例如，虚假信息常沿“低信誉用户→高信誉用户”的异常路径扩散，而真实信息的传播拓扑则更趋均匀。若跨模态交互过程忽视此类结构先验，模型将难以识别“图文表面一致但传播模式异常”的伪装型虚假信息。

为解决上述问题，本文提出结构引导的跨模态交互机制，如图2所示。其核心思想是：将传播拓扑信息以显式偏置的形式，注入跨模态注意力的未归一化注意力分数（logits）计算中，使传播结构作为先验项对图文交互强度进行约束与重加权[32-33]。

2.4.1 结构偏置的构建

为全面刻画传播结构对跨模态交互的引导作用，本文从四个维度对图进行解构和建模：传播影响力、传播时序、交互关系和拓扑结构。这四个维度分别对应节点重要性偏置、时序偏置、边类型偏置和节点度偏置，共同构成统一的结构偏置矩阵 $\boldsymbol{B}$：

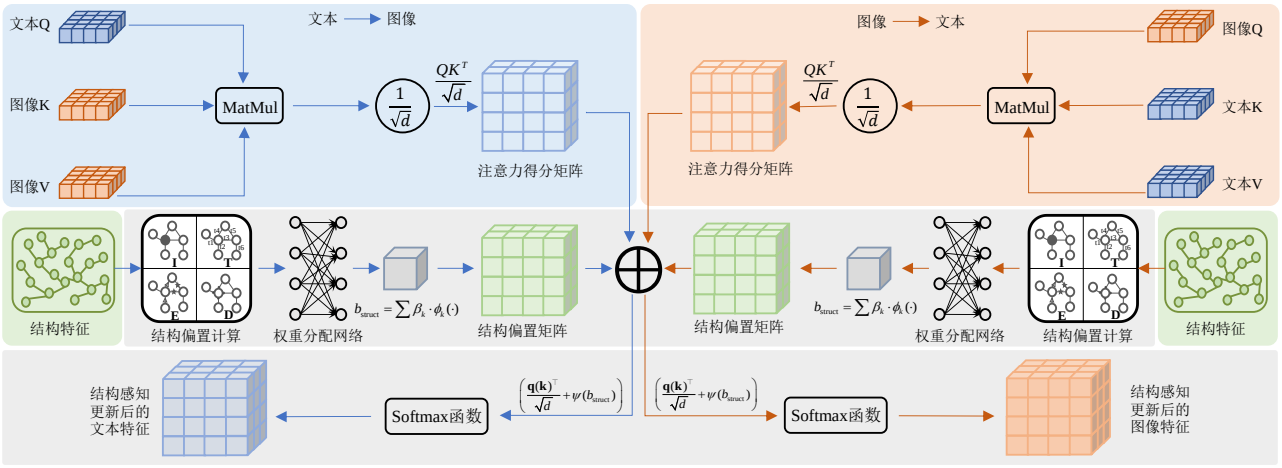


图2　结构感知跨模态注意力机制

（1）传播影响力偏置 $\boldsymbol{B}_{\text{imp}}$

在虚假信息传播过程中，不同用户节点的影响力存在显著异质性。为量化节点在传播拓扑中的结构重要性，本文引入结构传播影响力（Structural Propagation Influence，SPI）。设传播图为 $G=(U,E)$，其中 U 为节点集合，E 为边集合。有向边 $i \to j$ 代表用户 i 的信息被 j 转发。节点 i 的 SPI 值 s_i 通过递归方式定义：

$$s_i = \frac{1-\rho}{|\mathcal{U}|} + \rho \sum_{j \in \mathcal{N}_{in}(i)} \frac{s_j}{d_{out}(j)} \tag{8}$$

其中，$\mathcal{N}_{in}(i)$ 表示指向节点 i 的入边邻居集合（即在传播链中转发或引用 i 的用户），$d_{out}(j)$ 为节点 j 的出度，$|\mathcal{U}|$ 为传播图中的节点总数，$\rho \in (0,1)$ 为阻尼因子（本文取 $\rho = 0.85$）。该定义等价于对反向传播图进行中心性建模，即影响力沿信息传播路径的反方向进行累积：一个用户被越多高影响力用户转发，其 SPI 值越高。基于此，定义节点重要性偏置为 SPI 值的归一化均值：

$$\boldsymbol{B}_{\text{imp}}(i,j) = \frac{s_i + s_j}{2 \cdot \max_{u \in \mathcal{U}} s_u} \tag{9}$$

该偏置项通过强化影响力较大的节点对，使跨模态注意力能够聚焦于关键传播节点对应的跨模态交互。在本文设定下，SPI 的计算基于完整观测传播图，属于离线场景下的结构先验建模。

（2）传播时序偏置 $\boldsymbol{B}_{\text{time}}$

虚假信息扩散往往具有突发性，相邻转发之间的时间间隔极短。为捕捉这一动态特征，本文基于传播事件的时间戳构造时序偏置。对于传播图中的边 $i \to j$，记 t_i 和 t_j 分别为节点 i 和 j 参与传播的时间戳，定义时序偏置为：

$$\boldsymbol{B}_{\text{time}}(i,j) = \exp(-\lambda_t \cdot |t_i - t_j|) \tag{10}$$

其中 λ_t 为可学习的衰减系数（初始化为 0.1）。为保证模型训练的数值稳定性，避免 λ_t 为负导致指数项异常放大，实际实现时对 λ_t 施加 Softplus 激活函数约束其非负性。时间间隔越小，偏置值越接近 1，模型越倾向于强化该传播链路上的语义交互；反之，时间间隔越大，该偏置的引导作用越弱。

（3）交互关系偏置 $\boldsymbol{B}_{\text{edge}}$

社交平台上的用户交互行为包含转发、评论、引用等多种类型，不同类型的边在传播中的语义贡献存在差异。为此，本文为每种边类型分配一个可学习的标量权重，以区分其对跨模态对齐的贡献：

$$\boldsymbol{B}_{\text{edge}}(i,j) = w_{\text{edge}}, \quad w_{\text{edge}} \in \{w_{\text{retweet}}, w_{\text{reply}}, w_{\text{quote}}\} \tag{11}$$

其中 w_{retweet}、w_{reply}、w_{quote} 均为可训练参数，通过反向传播算法进行联合优化。

（4）拓扑结构偏置 $\boldsymbol{B}_{\text{deg}}$

节点的局部连接密集程度反映了其作为传播枢纽的潜力。对节点 i，定义其度偏置为归一化对数度：

$$\boldsymbol{B}_{\text{deg}}(i) = \frac{\log(d(i)+1)}{\max_{u \in \mathcal{U}} \log(d(u)+1)} \tag{12}$$

其中 $d(i)$ 为节点 i 的度，即入度与出度之和。为满足注意力机制的输入要求，采用对称扩展方式构建节点对偏置：

$$\boldsymbol{B}_{\text{deg}}(i,j) = \frac{\boldsymbol{B}_{\text{deg}}(i) + \boldsymbol{B}_{\text{deg}}(j)}{2} \tag{13}$$

2.4.2 偏置融合与注意力注入

上述四类偏置从不同维度刻画了传播结构的特性。为了实现多维拓扑特征的统一整合，本文引入轻量级门控融合机制：首先将四类偏置沿特征维度进行拼接，通过可学习的线性变换与Sigmoid激活函数生成融合权重，最后以加权求和方式得到综合结构偏置矩阵$\boldsymbol{B}$：

$$\boldsymbol{B}=\sigma\left(\boldsymbol{W}_g\cdot[\boldsymbol{B}_{\text{imp}};\boldsymbol{B}_{\text{time}};\boldsymbol{B}_{\text{edge}};\boldsymbol{B}_{\text{deg}}]+\boldsymbol{b}_g\right) \tag{14}$$

其中[;]代表通道维度的拼接。在注入阶段，本文选择将结构信息作为偏置项直接作用于Softmax的输入logits。相较于传统的特征叠加，这种方式具备较强的解释性：它在不改变模态原始语义空间的前提下，对跨模态注意力分布进行结构约束与重加权。此时，结构偏置矩阵与注意力得分矩阵具有相同的矩阵形状，可在Softmax之前逐元素相加，从而将传播结构先验融入注意力计算过程，对不同位置的注意力分配产生差异化影响。给定文本查询向量$\boldsymbol{Q}_t=\boldsymbol{H}_t\boldsymbol{W}_Q$与图像键向量$\boldsymbol{K}_v=\boldsymbol{H}_v\boldsymbol{W}_K$，修正后的注意力权重计算公式为：

$$\text{Attention}(\boldsymbol{Q}_t,\boldsymbol{K}_v,\boldsymbol{V}_v)=\text{softmax}\left(\frac{\boldsymbol{Q}_t\boldsymbol{K}_v^T}{\sqrt{d_k}}+\lambda\cdot\boldsymbol{B}\right)\boldsymbol{V}_v \tag{15}$$

其中$\boldsymbol{V}_v=\boldsymbol{H}_v\boldsymbol{W}_V$为图像值向量，$d_k$为缩放因子，$\lambda$为可学习缩放因子（初始化为0.1），可学习缩放因子起到了平衡结构偏置与语义相似度的作用，保障训练过程的平滑与稳健。语义相似度项用于刻画图文内容相关性，结构偏置项用于提供传播语境下的可信度先验。

2.4.3 文本和图像双向跨模态交互

为充分挖掘文本与图像之间的互补信息，本文采用双向跨模态注意力机制，同时计算文本到图像（$T\rightarrow V$）和图像到文本（$V\rightarrow T$）两个方向的注意力输出：

$$\boldsymbol{H}_{t\rightarrow v}=\text{CrossAttn}(\boldsymbol{Q}=\boldsymbol{H}_t,\boldsymbol{K}=\boldsymbol{H}_v,\boldsymbol{V}=\boldsymbol{H}_v,\boldsymbol{B}) \tag{16}$$

$$\boldsymbol{H}_{v\rightarrow t}=\text{CrossAttn}(\boldsymbol{Q}=\boldsymbol{H}_v,\boldsymbol{K}=\boldsymbol{H}_t,\boldsymbol{V}=\boldsymbol{H}_t,\boldsymbol{B}^T) \tag{17}$$

其中B^T为结构偏置矩阵的转置。将两个方向的增强表示进行拼接，并通过线性投影层融合，得到结构增强的跨模态联合表示：

$$\boldsymbol{H}_c=\boldsymbol{W}_c\cdot[\boldsymbol{H}_{t\rightarrow v};\boldsymbol{H}_{v\rightarrow t}]+\boldsymbol{b}_c \tag{18}$$

其中$\boldsymbol{H}_c\in\mathbb{R}^d$为融合了传播结构引导信息的跨模态语义向量。该表示同时保留了多模态语义的一致性特征与传播结构的先验约束，为后续的自适应融合提供了高判别力的特征基础。

2.5 自适应多模态融合与分类

在真实社交媒体环境中，不同样本的模态质量存在显著差异。部分样本可能包含大量文本噪声或传播图极度稀疏。若采用固定权重进行融合，低质量模态的噪声将被放大，严重干扰模型的判别鲁棒性。针对上述问题，本文提出一种样本级自适应多模态融合方法。根据当前样本各模态的特征分布，动态评估其判别价值，并生成模态权重系数。将跨模态交互输出的联合表示$\boldsymbol{H}_c$与图编码表示$\boldsymbol{H}_g$拼接，构建融合上下文向量$\boldsymbol{H}_{\text{concat}}=[\boldsymbol{H}_c\|\boldsymbol{H}_g]\in\mathbb{R}^{512}$。该向量整合了图文联合语义与纯拓扑信息，将其输入由两层全连接层构成的轻量级门控网络：

$$\boldsymbol{h}_{\text{gate}}=\text{ReLU}(\boldsymbol{W}_1\cdot\boldsymbol{H}_{\text{concat}}+\boldsymbol{b}_1) \tag{19}$$

$$[\alpha_t,\alpha_v,\alpha_g]=\text{softmax}(\boldsymbol{W}_2\cdot\boldsymbol{h}_{\text{gate}}+\boldsymbol{b}_2) \tag{20}$$

其中$\boldsymbol{W}_1\in\mathbb{R}^{256\times512}$，$\boldsymbol{W}_2\in\mathbb{R}^{3\times256}$。门控网络输出的$\alpha_t,\alpha_v,\alpha_g$分别为文本、图像及传播结构模态的动态权重。当某模态特征存在严重噪声时，门控网络将自动降低其权重。通过模态特定的线性投影层对各表示进行空间对齐后，以加权求和方式获得最终的多模态融合表示：

$$\boldsymbol{H}_{\text{final}}=\alpha_t\cdot(\boldsymbol{W}_{t'}\boldsymbol{H}_t)+\alpha_v\cdot(\boldsymbol{W}_{v'}\boldsymbol{H}_v)+\alpha_g\cdot(\boldsymbol{W}_{g'}\boldsymbol{H}_g) \tag{21}$$

其中$\boldsymbol{W}_{t'},\boldsymbol{W}_{v'},\boldsymbol{W}_{g'}\in\mathbb{R}^{256\times256}$。最后，将融合表示$\boldsymbol{H}_{\text{final}}$送入单层全连接网络与Sigmoid激活函数构成的分类器进行预测：

$$\hat{y}_i=\text{Sigmoid}(\boldsymbol{W}_{\text{out}}\cdot\boldsymbol{H}_{\text{final}}+\boldsymbol{b}_{\text{out}}) \tag{22}$$

模型训练采用二分类交叉熵损失函数：

$$\mathcal{L}=-\frac{1}{M}\sum_{i=1}^{M}\left[y_i\log\hat{y}_i+(1-y_i)\log(1-\hat{y}_i)\right] \tag{23}$$

通过端到端优化，自适应融合机制使模型在面对数据噪声干扰等非理想条件时，具有更好的适应能力和稳定性。

3 实验结果与分析

为全面验证所提SGMAF模型的有效性与鲁棒性，本文围绕以下几个方面开展实验分析：

（1）验证SGMAF在多模态虚假信息检测任务中的整体性能，并与现有主流方法进行对比；

（2）分析结构引导机制与样本级自适应多模态融合方法对模型性能的具体贡献；

（3）通过跨模态语义对齐可视化分析，从表征空间层面解释结构引导机制对跨模态交互过程的影响，并进一步揭示该对齐过程对样本级自适应融合策略的作用基础；

（4）评估关键超参数对模型性能的影响以及模型在不同场景下的稳定性；

（5）评估SGMAF模型在实际复杂场景下的检测行为与决策依据。

3.1　实验设置

3.1.1　实验数据集

本研究引入Pheme[34]与微博[35]公开数据集，以量化SGMAF框架的判别性能。Pheme语料库源自推特，其内嵌的密集交互记录，在现有文献中被认为能为传播拓扑分析提供坚实支撑。微博基准则来源于中文网络，该数据集携带了更为丰富的视觉表征，以此校验模型在高维多模态场景下的鲁棒性边界。原始数据的残缺极易干扰特征映射，清洗阶段删除了缺失视觉或文本成分的样本，确立三元组，所有进入计算空间的实例均具备完整的“文本-图像-传播图”结构，重构后的基准统计细节详见表1。

表1　数据集样本分布与传播结构统计对比

数据集类型	虚假信息样本	真实信息样本	虚假信息比例	传播结构节点数	传播结构边数
Pheme	590	1428	29.24%	4438	2191
微博	590	877	40.22%	6963	3080

3.1.2　评估指标

虚假信息检测本质上属于二分类任务。为全面评估模型性能，本文选取准确率（Accuracy）、精确率（Precision）[36]、召回率（Recall）及F1值（F1-score）作为主要评价指标[37]。Accuracy衡量模型整体分类的正确率；Precision与Recall则分别反映模型识别虚假信息时的精准程度与覆盖能力。F1-score作为两者的调和平均值，通常被认为能更稳健地反映类别不平衡条件下的整体性能。上述指标的计算公式如下：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{24}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{25}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{26}$$

$$\text{F1 - score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{27}$$

其中，TP（True Positive）代表被模型正确识别为虚假的样本数目；TN（True Negative）指被正确判断为真实的样本数目。FP（False Positive）对应真实样本被误标为虚假的情况，而FN（False Negative）则表示虚假样本被误判为真实的数量。在现有评估框架下，综合这些指标，能够较为全面地衡量模型在虚假信息检测中的分类能力。

3.1.3　参数设置

所有实验均基于PyTorch框架实现，并在配备NVIDIA RTX 4090 GPU的服务器环境下完成。模型采用Adam优化器和早停策略进行训练。为保证实验的稳定性与可复现性，统一设置超参数如下：最大训练轮数为20，批处理大小为32，文本序列最大长度为50，跨模态注意力头数为8，Dropout率为0.6，初始学习率设定为$2e-3$。

3.2　整体性能评估

为验证所提出模型SGMAF在多模态虚假信息检测任务中的有效性，本文选取了近年来具有代表性的多模态虚假信息检测方法作为对比模型，涵盖基于生成式建模、图神经网络、多模态深度融合以及大规模多模态模型等方法。具体包括MVAE、SAFE、EBGCN、MFAN、NLIN、CLFFRD、MIMoE-FND以及Human-Like等方法。其中，MVAE通过变分自编码器对多模态数据进行联合潜在空间建模，实现跨模态信息融合；SAFE基于图文语义一致性建模，从内容相似性角度刻画文本与图像之间的关联；EBGCN将图结构引入传播建模，并结合贝叶斯推断增强传播关系表达能力；MFAN利用跨模态注意力机制对文本、图像及传播结构信息进行自适应融合；NLIN将多模态信息映射至统一语言推理空间，通过上下文推理完成虚假信息判别；CLFFRD则引入课程学习策略，结合细粒度跨模态融合机制逐步提升模型表达能力。MIMoE-FND采用混合专家机制，对模态交互进行专长分工与动态路由；Human-Like采用大模型与小模型协同框架，采用Qwen-VL-Max完成图文联合推理，并利用Qwen-Turbo生成反思式摘要，将高层语义线索与

轻量级融合网络结合，实现基于大规模多模态模型的推理增强检测。

实验结果如表 2 和表 3 所示，MVAE 在两个数据集上表现均不理想，Pheme 准确率 78.26%，微博 72.18%。这表明单纯依赖变分自编码器学习跨模态共享分布，在语义高度复杂的虚假信息面前，因缺少显式的语义对齐约束和拓扑特征支撑，难以捕获强判别性异质特征，泛化能力受限。

SAFE 通过度量文本与图像的语义相似度来判别矛盾。在微博数据集上，其准确率较 MVAE 提升 13.24 个百分点，这一幅度表明，模态间语义一致性建模可成为检测多模态虚假信息的有效出发点。但 SAFE 忽略了社交网络传播动态，在视觉线索匮乏的 Pheme 数据集上，性能增益不及微博显著。

EBGCN 专注传播结构的贝叶斯图建模，在 Pheme 上优于内容驱动的基线方法。这说明在传播链条较完整的推特环境中，拓扑信息是甄别虚假信息的核心线索。然而，在多模态信息密集的微博上，其准确率落后于多模态融合模型，反映出单一结构建模难以应对图文混杂场景。

MFAN 整合文本、图像与图结构，两个数据集 F1 分数均超 85%。但 MFAN 仅实现模态间平行交互，缺乏结构信息对图文对齐的显式引导。作为对比，SGMAF 通过结构引导跨模态交互机制，图结构有效引导图文交互，在微博上将准确率从 88.81% 进一步推至 96.27%。

表 2 模型在 Pheme 数据集中的对比实验结果（%）

模型	准确率	精确率	召回率	F1 分数
MVAE	78.26	74.30	73.15	73.58
SAFE	82.29	81.06	80.27	80.34
EBGCN	84.41	82.39	78.62	80.12
MFAN	88.31	86.23	85.26	85.72
NLIN	90.30	87.50	88.30	87.90
CLFFRD	89.95	88.26	87.57	88.13
MIMoE-FND	88.05	85.24	86.62	85.87
Human-Like	89.87	89.73	**89.86**	**89.74**
SGMAF	**90.91**	**90.28**	87.36	88.64

NLIN 和 CLFFRD 代表当前语义推理的先进水平。NLIN 将多模态信息转化为统一推理逻辑，在微博上取得 92.20% 准确率；CLFFRD 利用课程学习强化细粒度融合。但这两类方法未考虑样本间模态质量差异，也缺乏对传播路径异质影响力的刻画。

表 3 模型在微博数据集中的对比实验结果（%）

模型	准确率	精确率	召回率	F1 分数
MVAE	72.18	71.03	70.72	70.85
SAFE	85.42	84.65	84.95	84.79
EBGCN	82.71	82.41	83.93	82.45
MFAN	88.81	89.16	87.30	88.02
NLIN	92.20	91.70	92.20	91.90
CLFFRD	91.26	90.23	88.70	89.92
MIMoE-FND	92.60	93.36	92.54	92.37
Human-Like	89.50	88.83	89.37	89.07
SGMAF	**96.27**	**96.34**	**96.27**	**96.24**

MIMoE-FND 在两个数据集上较 MFAN、CLFFRD 等均有提升，表明 MoE 机制能够增强模态交互的表达能力，但其整体表现仍不及 SGMAF，说明仅靠模态专家划分不足以刻画传播结构与内容语义间的耦合关系。基于大规模多模态模型的检测基线 Human-Like 在 Pheme 数据集上表现突出，尤其在召回率与 F1 值方面具有一定优势，表明其在图文内容语义建模方面具有较强表现。SGMAF 在准确率与精确率方面仍保持领先，说明传播拓扑结构信息能够有效补充内容语义信息，从而提升模型对复杂样本的判别能力。与 Human-Like 侧重图文内容语义不同，本文进一步融合传播拓扑结构信息，形成“结构+内容”的联合建模范式，两类方法在信息利用维度上具有一定的互补性。在微博数据集上，SGMAF 四项指标均取得最优。

在 Pheme 数据集中，本文提出的 SGMAF 模型结构引导机制有效建模传播拓扑差异，精确率达 90.28%。在更复杂的微博场景下，SGMAF 凭借样本级自适应融合机制，动态放大高质量视觉模态的判别权重，准确率领先次优模型 4 个百分点。综合结果表明，SGMAF 通过结构引导与自适应调节的协同，在不同社交平台的传播生态中均展现出较强的适应性。

3.3 模块消融实验分析

为了进一步剖析 SGMAF 模型中各个关键模块在虚假信息检测任务中所发挥的具体作用，本文在

Pheme与微博两个真实数据集上分别构建了五组消融实验变体，并对其性能进行了对比分析。各消融项与2.2节所述核心部分的对应关系如下：-w/o B为移除结构引导跨模态交互中的结构偏置注入；-w/o A为移除样本级自适应多模态融合；-w/o BA为同时移除两者；-w/o G为移除多模态特征编码模块中的图结构模态；-w/o V为移除视觉模态。相关实验结果如表4（Pheme）和表5（微博）所示。具体设定如下：

（1）-w/o B：移除结构偏置，跨模态交互退化为仅依赖图文内容的常规注意力机制；

（2）-w/o A：移除自适应多模态融合，采用固定权重对各模态特征进行直接拼接与映射；

（3）-w/o BA：同时移除结构偏置与自适应多模态融合，作为基础的多模态图神经网络；

（4）-w/o G：移除图结构模态，仅保留文本与图像作为模型输入；

（5）-w/o V：移除视觉模态，仅保留文本与图结构作为模型输入。

表4　Pheme数据集的消融实验结果（%）

模型	准确率	精确率	召回率	F1分数
SGMAF	90.91	90.28	87.36	88.64
-w/o B	83.38	80.44	78.41	79.29
-w/o A	84.94	84.28	77.96	80.15
-w/o BA	83.11	81.96	75.37	77.53
-w/o G	85.71	82.45	84.46	83.31
-w/o V	86.49	86.27	80.10	82.23

表5　微博数据集的消融实验结果（%）

模型	准确率	精确率	召回率	F1分数
SGMAF	96.27	96.34	96.27	96.24
-w/o B	91.52	90.84	91.80	91.23
-w/o A	92.88	94.75	90.94	92.25
-w/o BA	89.49	88.78	90.12	89.20
-w/o G	93.89	95.43	92.24	93.40
-w/o V	88.81	89.15	87.29	88.01

在Pheme与微博两个数据集上，完整模型SGMAF均取得最优性能。同时移除结构偏置与自适应融合（-w/o BA）造成的性能退化最为严重，Pheme F1降至77.53%，微博F1降至89.20%，降幅显著超出任一单模块移除的情形。结构偏置模块的独立贡献在Pheme上尤为突出：移除后（-w/o B）精确率从90.28%跌至80.44%，在微博上同样引发明显退化，表明传播拓扑先验对跨模态注意力聚焦关键区域的作用具有跨平台稳定性。自适应融合模块（-w/o A）的移除在两个数据集上均导致F1下降，固定融合策略难以弥补样本级动态权重调节的缺失。

模态贡献的分布随数据集不同而显著分化。Pheme数据集中，图结构模态（-w/o G）的移除使精确率从90.28%降至82.45%，跌幅领先于其他模态，传播结构信息在降低误报方面承担主导角色。微博数据集则呈现相反的趋势：视觉模态（-w/o V）移除后F1从96.24%骤降至88.01%，降幅8.23个百分点，损失远超其他消融项；图像中携带的图文不一致等判别线索，在此场景下构成了模型的主要倚仗。图结构模态移除后微博F1仍维持93.40%，其作用更多体现为对模型输出的辅助稳定，而非核心判别依据。结构偏置与自适应融合并非两个独立的增益来源。前者从传播拓扑层面施加全局约束，后者在样本层面实现动态调节。二者的协同使得模型能够根据数据分布的偏移，在不同模态之间重新分配依赖权重。这种功能互补关系，而非单一模块的绝对贡献，构成了SGMAF跨平台性能提升的底层逻辑。

3.4　模态缺失鲁棒性分析

为进一步评估所提模型在图文输入不完整条件下的稳定性，本文在测试阶段构造随机模态缺失场景，分别模拟缺图和缺文本两类情况。实验过程中保持训练过程与模型参数不变，仅对指定模态特征进行随机置零，以观察模型在不完整输入下的性能变化趋势。该实验的目的在于补充分析模型在图文模态缺失条件下的鲁棒性边界，而不改变本文完整“文本—图像—传播图”三元组的主任务设定。

实验中，缺失率设为$p\in\{0.0,0.2,0.4,0.6\}$，其中$p=0.0$为完整输入条件。为保证实验设定的一致性，仅选取能够自然适配相同缺失协议的代表性模型进行对比，重点比较SGMAF、SGMAF-w/o A和MFAN。评价指标仍采用准确率和F1分数。

如表6和表7所示，在缺图场景下，SGMAF在两个数据集上的性能下降相对平缓。关于F1分数，Pheme数据集从88.64%降至84.60%，微博从

96.24%降至89.25%。相比之下，MFAN的下降幅度更大，说明在图像缺失时，SGMAF对模态缺失的适应性更强。在缺文本场景下，各模型性能同样随缺失率增大而下降。SGMAF仍保持相对优势，以Pheme数据集为例，SGMAF与MFAN两者F1分数的差距从完整输入时的2.92个百分点扩大至缺失率0.6时的6.30个百分点。这表明，即使在文本信息大量缺失时，样本级自适应融合机制仍能有效利用其他模态，抑制性能退化。

综上所述，随着缺失率增加，三种模型在缺图和缺文本场景下的性能均呈下降趋势，但SGMAF的退化幅度整体小于MFAN和SGMAF-w/o A。结果表明，样本级自适应多模态融合方法在图文模态质量波动时具有一定的缓冲作用，能够在一定程度上减弱模态缺失对最终判别的影响。本文方法在完整输入条件下有效，同时在不完整输入条件下也具有一定的稳定性。

3.5 跨模态语义对齐可视化分析

本节以Pheme数据集为例，从模型训练过程中选取六个关键状态，对表征变化趋势进行可视化分析，以探究结构引导机制在跨模态交互中的作用及其对特征对齐效果的影响。

图3展示了各模态特征空间分布与跨模态余弦相似度分布的动态变化。散点图中，圆形、正方形和三角形分别对应文本、图像与传播图特征。跨模态交互模块输出各模态的高维表征后，经由t-SNE降维完成可视化。观察二维空间中模态分布的演变，可直观分析跨模态对齐过程及传播结构的约束作用；同时计算“文本-图像”“文本-传播图”“图像-传播图”三组余弦相似度并绘制分布直方图，从统计层面刻画语义一致性的变化趋势。如图3所示，训练初期如子图a和b所示，各模态特征呈明显分离，语义鸿沟显著。随训练推进，结构引导机制使文本与图像特征向共享空间逐步收敛，并在局部区域建立交互关系；传播图特征则始终保持相对独立的流形结构，对跨模态对齐施加约束。至收敛阶段。如子图f所示，图文特征呈现清晰的局部聚集趋势，传播图仍维持结构特异性。从相似度分布看，“文本-图像”相似度逐步提升并稳定于0.8附近，反映图文语义一致性显著增强；“文本/图像-传播图”相似度则稳定在0.4～0.5区间，表明结构信息与内容语义之间保留了适度的异质性，未发生表征塌缩。

上述结果表明，SGMAF的结构引导机制在增强跨模态语义一致性的同时，有效保留了模态间的结构差异，实现了一种兼顾对齐与异构性的联合表征学习。这一特性为后续样本级自适应融合提供了稳定且具有判别力的表示基础。

表6 缺图场景下不同缺失率的性能对比（%）

	Pheme						微博					
	SGMAF		SGMAF-w/o A		MFAN		SGMAF		SGMAF-w/o A		MFAN	
缺失率	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数
0.0	90.91	88.64	84.94	80.15	88.31	85.72	96.27	96.24	92.88	92.25	88.81	88.02
0.2	89.85	87.52	83.12	78.40	86.85	84.15	94.85	94.62	90.15	89.60	86.20	85.35
0.4	88.40	86.15	80.75	76.10	84.45	81.55	92.65	92.35	86.40	85.75	82.75	81.80
0.6	86.82	84.60	77.92	73.25	81.30	78.42	89.75	89.25	82.15	81.30	78.60	77.55

表7 缺文本场景下不同缺失率的性能对比（%）

	Pheme						微博					
	SGMAF		SGMAF-w/o A		MFAN		SGMAF		SGMAF-w/o A		MFAN	
缺失率	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数
0.0	90.91	88.64	84.94	80.15	88.31	85.72	96.27	96.24	92.88	92.25	88.81	88.02
0.2	89.15	86.75	82.10	77.20	85.60	82.95	95.12	95.05	90.95	90.25	87.05	86.15
0.4	86.60	84.20	78.45	73.40	81.90	79.10	93.45	93.25	88.10	87.35	84.60	83.60
0.6	83.45	80.95	74.20	69.15	77.50	74.65	91.10	90.85	84.65	83.80	81.25	80.15

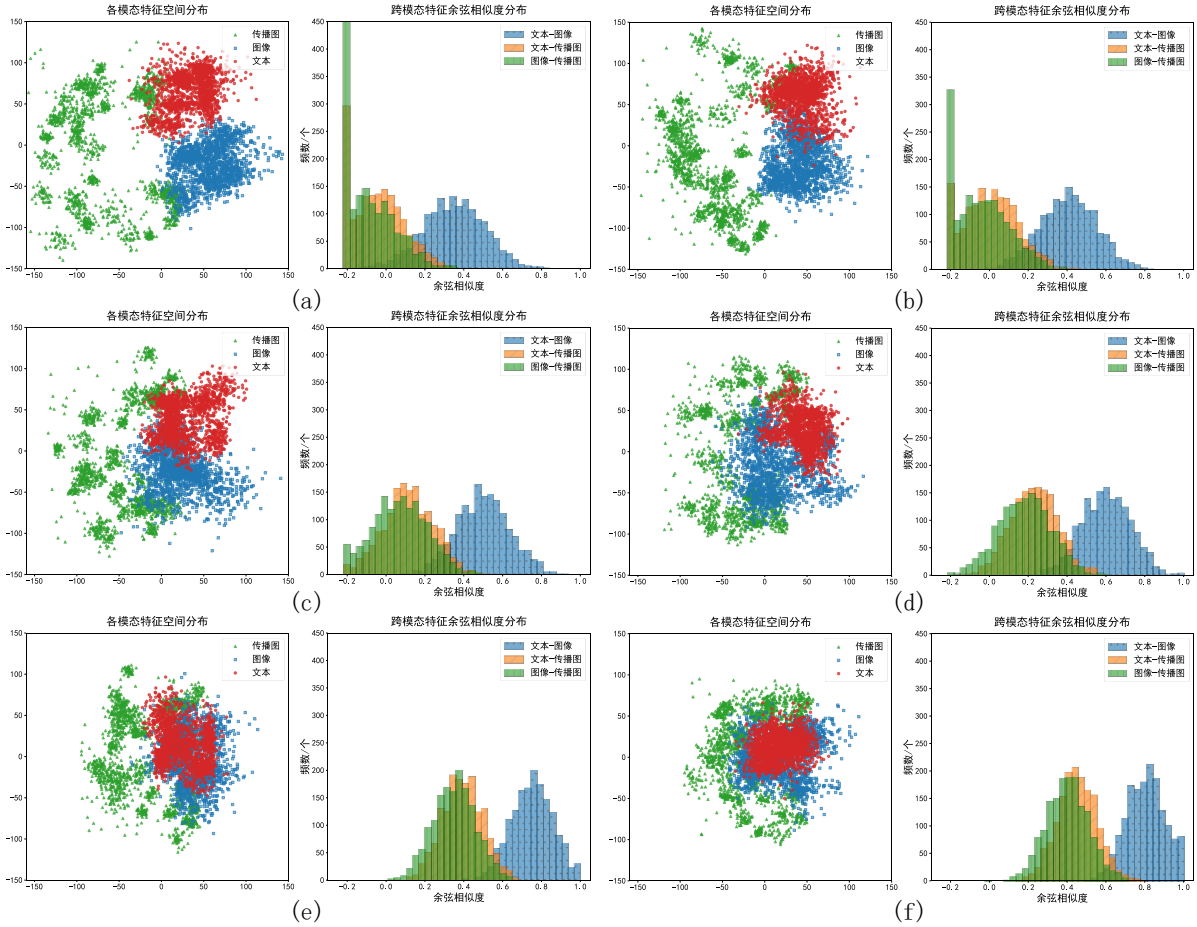


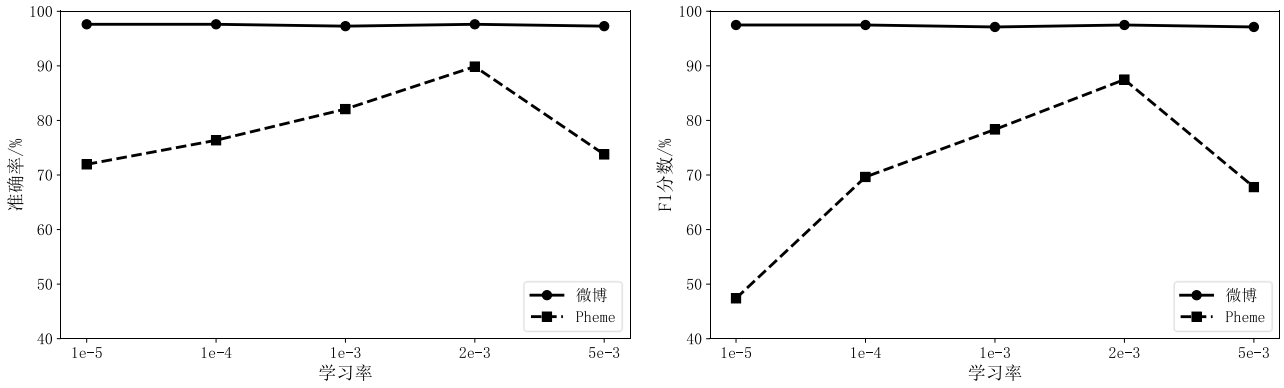
图3　结构引导下的跨模态语义对齐与特征空间演化过程

3.6　参数敏感性分析

为评估关键超参数对模型性能的影响，本文选取学习率与随机失活率，在Pheme与微博两个数据集上进行敏感性分析，实验结果如图4和图5所示。学习率控制参数更新步长，影响模型的收敛速度与优化稳定性。本文在{1*e* - 5,1*e* - 4,1*e* - 3,2*e* - 3,5*e* - 3}范围内进行实验。图4显示，两个数据集上的性能均随学习率增大呈先升后降的趋势。学习率过低时，参数更新步长偏小，收敛缓慢，有限训练轮数内难以逼近最优解，导致整体性能受限。随学习率提升至适中水平，参数更新效率改善，性能同步增长。当学习率取值为2*e* - 3时，两个数据集均达到最优，表明该取值区间内收敛效率与训练平稳性实现了较好平衡。但学习率进一步增大至5*e* - 3时，性能出现明显下滑。其原因是步幅过大导致优化过程产生震荡，甚至偏离最优区域，最终损害模型性能。

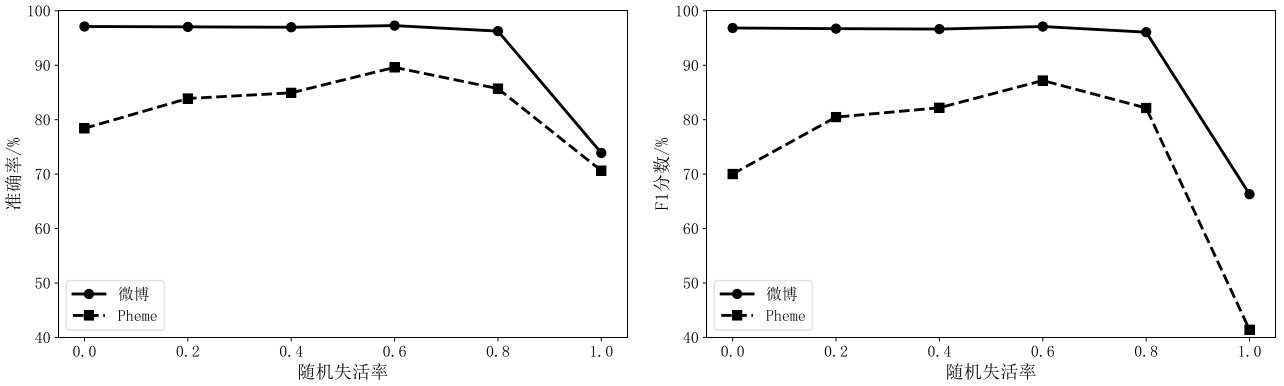
Dropout是常用的正则化手段，通过随机丢弃神经元来缓解过拟合。本文在{0.2,0.4,0.6,0.8,1.0}范围内对该参数进行分析。图5显示，两个数据集上的性能随Dropout比率增大同样呈先升后降的趋势。比率为0.2时，正则化作用偏弱，模型易对训练数据产生过拟合，泛化能力受限，整体性能处于较低水平。随比率逐步增大，模型对冗余特征的依赖减弱，能够学习到更稳健的表征，性能随之提升。当比率取0.6时，两个数据集均达到最优，表明适度强度的神经元随机丢弃有利于增强泛化能力与鲁棒性。但比率继续上调至0.8或更高后，检测效果转入下滑通道；尤其取1.0时性能退化明显。其原因是失活率过高会屏蔽大量特征信息，破坏表示学习的完整性，使模型难以有效捕捉多模态关联，进而导致欠拟合。

综合学习率与Dropout的分析，两个参数对模型性能均影响显著，且在不同数据集上的影响规律基本一致。参数取值需落在合理区间内，方能在收敛速度与泛化能力之间取得合意折中。基于上述实



(a)准确率随学习率变化曲线　　(b)F1分数随学习率变化曲线

图4　学习率对实验结果的影响



(a)准确率随失活率变化曲线　　(b)F1分数随失活率变化曲线

图5　随机失活率对实验结果的影响

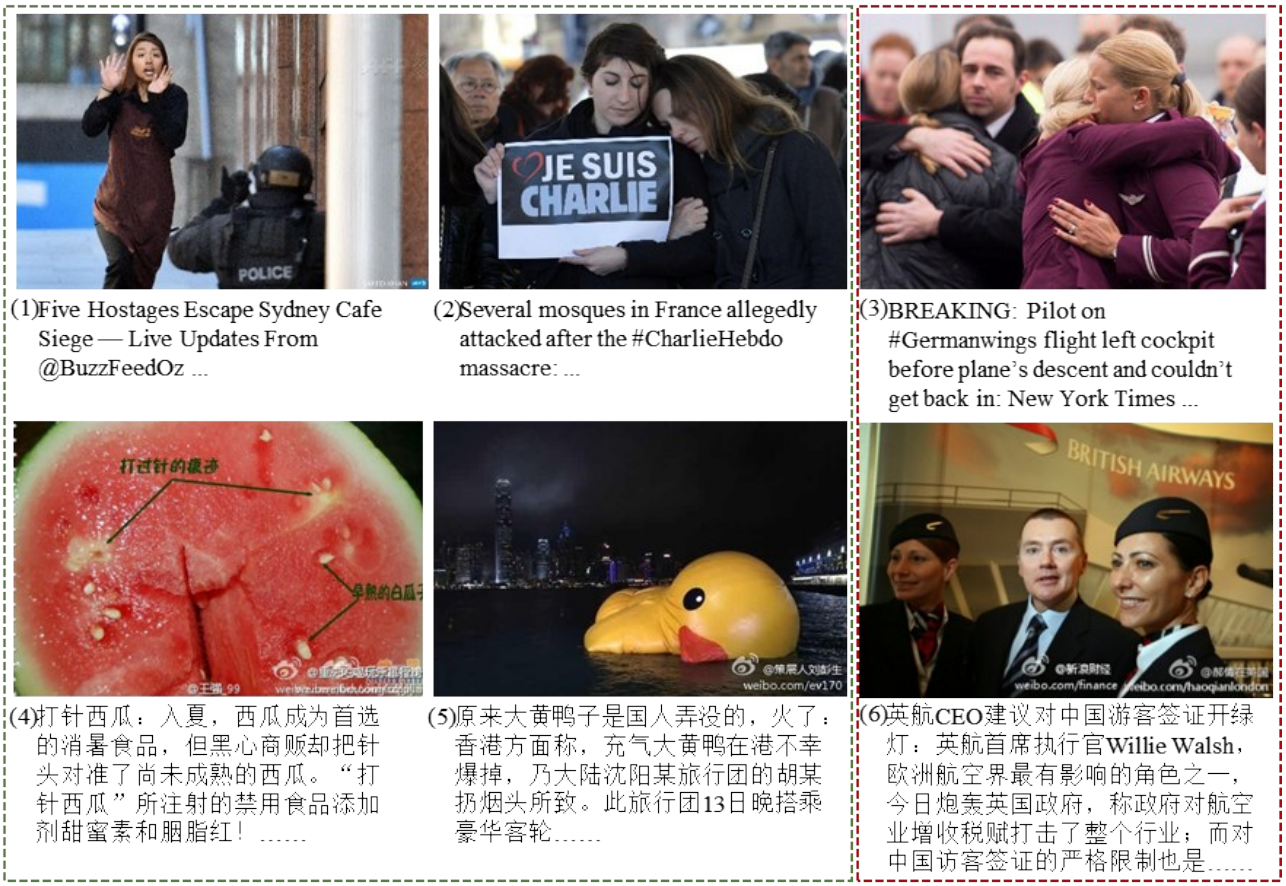
验，本文实验将学习率固定为$2e-3$，Dropout比率设为0.6。

3.7　案例分析

为进一步评估SGMAF模型在复杂场景下的检测行为与判别依据，本文从两个数据集中选取典型样例进行定性分析，案例示例如图6所示。其中，正确分类的样本用绿色虚线框住，错误分类的样本用红色虚线框住。为便于对比分析，表8进一步总结了不同案例的信息类型、多模态表征特征及模型行为。

Pheme数据集主要包含突发事件场景下的传播树结构与事件相关图文信息，更侧重考察模型对传播拓扑模式的建模能力。从表8与图6可以看出，案例1中的虚假信息在传播过程中呈现浅层、高密度的扩散特征，SGMAF借助结构引导机制有效捕获了异常传播模式，实现了正确识别；案例2则表现出较为规范的层级化传播规律，模型能够较好地学习真实新闻的有序扩散特征。案例3中的虚假信息引用权威媒体内容并采用规范化新闻语体，使图文语义呈现较高一致性，其传播路径也与真实新闻较为接近。在该场景下，结构引导机制对高影响力节点赋予了较强偏置，弱化了潜在异常传播信号；同时，较强的跨模态一致性进一步增强了模型对其真实性的倾向判断，最终导致误判。该结果表明，SGMAF更擅长识别传播异常型虚假信息，而对于结构与语义均高度合规的深度伪装场景仍存在一定局限。

微博数据集中，案例4的文本与图像存在明显语义冲突，模型能够捕捉文本情绪极化与视觉异常之间的不一致关系，实现了对虚假信息的正确识别。案例5呈现典型的群体情绪驱动扩散特征，传播结构与多模态语义形成互补验证，进一步增强了模型在复杂场景下的鲁棒性。但案例6中，虚假信息在文本表达、视觉内容与传播行为三个层面均高



(1)Five Hostages Escape Sydney Cafe Siege — Live Updates From @BuzzFeedOz ...

(2)Several mosques in France allegedly attacked after the #CharlieHebdo massacre: ...

(3)BREAKING: Pilot on #Germanwings flight left cockpit before plane's descent and couldn't get back in: New York Times ...

(4)打针西瓜：入夏，西瓜成为首选的消暑食品，但黑心商贩却把针头对准了尚未成熟的西瓜。“打针西瓜”所注射的禁用食品添加剂甜蜜素和胭脂红！ ……

(5)原来大黄鸭子是国人弄没的，火了：香港方面称，充气大黄鸭在港不幸爆掉，乃大陆沈阳某旅行团的胡某扔烟头所致。此旅行团13日晚搭乘豪华客轮……

(6)英航CEO建议对中国游客签证开绿灯：英航首席执行官Willie Walsh，欧洲航空界最有影响的角色之一，今日炮轰英国政府，称政府对航空业增收税赋打击了整个行业；而对中国访客签证的严格限制也是……

图6　SGMAF 模型两类数据集上的检测案例

度模仿真实新闻，形成了较强的“跨模态一致性”。在此情况下，样本级自适应多模态融合方法倾向于强化一致性较高的模态特征，而传播结构中缺乏明显异常模式，也进一步削弱了模型对潜在虚假属性的感知能力。由于 SGMAF 主要依赖传播结构与跨模态语义关联进行判别，因此在深度伪装场景下仍存在一定误判风险。

综合两个数据集的案例分析可以看出，SGMAF 在传播结构建模与跨模态语义融合方面具有较好的协同能力。结构引导机制能够有效捕获异常

表 8　SGMAF 模型典型案例的检测行为与特征分析

数据集	案例编号	信息类型	多模态表征特征	模型预测	模型行为分析
Pheme	1	突发事件型虚假信息	事件文本与关联配图共同传播	虚假信息	结构引导方法有效捕获异常扩散模式
Pheme	2	权威媒体真实报道信息	图文表征整体规范	真实信息	模型能够学习真实新闻的有序传播规律
Pheme	3	权威源伪装型虚假信息	常规媒体配图与权威文本表述形成“合规性伪装”	真实信息	高信誉源与规范化表征导致结构趋同，模型出现误判
微博	4	图文冲突型虚假信息	图像异常与文本情绪极化存在明显语义冲突	虚假信息	多模态融合方法有效捕捉跨模态不一致性
微博	5	群体情绪驱动型虚假信息	情绪化文本与事件图像形成强化传播	虚假信息	图结构与情绪语义形成互补验证
微博	6	高一致性伪装型虚假信息	图文语义高度一致；新闻语体规范	真实信息	跨模态一致性削弱异常特征，模型缺乏细粒度事实校验能力

扩散模式，样本级自适应多模态融合方法则能够增强模型对图文冲突与模态质量波动场景的鲁棒性。然而，当虚假信息在传播结构、文本语义与视觉内容三个层面均高度模仿真实新闻时，模型仍可能受到“结构合规性”与“跨模态一致性”的共同误导。该结果说明，SGMAF 主要提升的是复杂传播场景下的鲁棒检测能力，而对于缺乏显式异常信号的深度伪装型虚假信息仍存在一定局限。

4 结束语

聚焦社交媒体虚假信息检测中跨模态交互缺乏传播结构约束、固定融合方式难以适应模态质量差异的问题，本文提出了结构引导的自适应多模态融合模型 SGMAF。模型从传播网络中提取节点重要性、传播时序、边类型及节点度等拓扑信息，构建结构偏置矩阵并融入跨模态注意力计算，使传播行为对图文语义交互施加显式约束；同时提出样本级自适应多模态融合方法，根据各模态在判别任务中的实际贡献动态分配融合权重，以抑制低质量模态与噪声信息的干扰。在 Pheme 与微博两个真实数据集上的实验表明，SGMAF 在准确率、F1 值等多项指标上均优于当前主流基线模型。消融实验验证了结构偏置与自适应融合方法的协同增益。案例分析则从具体样本层面展示了模型在捕获异常传播模式与跨模态语义冲突方面的有效性。然而，案例分析结果也表明，当虚假信息在传播拓扑、文本语义与视觉内容三个层面均高度模仿真实新闻时，模型仍存在一定误判风险。未来将进一步结合外部知识增强与实体级事实校验机制，以提升模型在复杂真实性推断场景中的泛化能力。

参考文献：

[1] Abdali S, Shaham S, Krishnamachari B. Multi-modal misinformation detection: Approaches, challenges and opportunities[J]. ACM Computing Surveys, 2024, 57(3): 1-29.

[2] Mostafa M F, Almogren A S, Al-Qurishi M, et al. Modality deep-learning frameworks for fake news detection on social networks: A systematic literature review[J]. ACM Computing Surveys, 2024, 57(3): 1-50.

[3] Park S, Nan X. Generative AI and misinformation: a scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation[J]. AI & Society, 2026, 41(2): 1501-1515.

[4] Wu F, Jin H, Chen F, et al. Balanced cross-modal prompt learning and fusion network for multi-modal fake news detection[J]. Information Fusion, 2025, 122: 103196.

[5] Silva A, Luo L, Karunasekera S, et al. Unsupervised domain-agnostic fake news detection using multi-modal weak signals[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 7283-7295.

[6] Guo Y, Zhen K, Liu J. Contrastive Prompt Learning in Structured Graph Networks for Multimodal Fake News Detection[J]. IEEE Transactions on Big Data, 2026, 12(3): 1045-1057.

[7] Zhou Y, Yang Y, Ying Q, et al. Multimodal fake news detection via clip-guided learning[C]//2023 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2023: 2825-2830.

[8] Xu F, Zeng L, Huang Q, et al. Hierarchical graph attention networks for multi-modal rumor detection on social media[J]. Neurocomputing, 2024, 569: 127112.

[9] Li X, Qiao J, Yin S, et al. A Survey of Multimodal Fake News Detection: A Cross-Modal Interaction Perspective[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2025, 9(4): 2658-2675.

[10] Lv J, Gao Y, Li L, et al. Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges [J]. Journal of King Saud University Computer and Information Sciences, 2025, 37(9): 306.

[11] Luvembe A M, Li W, Li S, et al. An adaptive auto fusion with hierarchical attention for multimodal fake news detection[J]. Expert Systems with Applications, 2025, 285: 127930.

[12] Zhou X, Zafarani R. A survey of fake news: Fundamental theories, detection methods, and opportunities[J]. ACM Computing Surveys, 2020, 53(5): 1-40.

[13] Capuano N, Fenza G, Loia V, et al. Content-based fake news detection with machine and deep learning: a systematic review[J]. Neurocomputing, 2023, 530: 91-103.

[14] Zhang X, Cao J, Li X, et al. Mining Dual Emotion for Fake News Detection[C]//Proceedings of The Web Conference 2021. New York: ACM Press, 2021: 3465-3476.

[15] 柯婧, 谢哲勇, 徐童, 陈宇豪, 廖祥文, 陈恩红. 基于大语言模型隐含语义增强的细粒度虚假新闻检测方法[J]. 计算机研究与发展, 2024, 61(5): 1250-1260.
Ke J, Xie Z Y, Xu T, Chen Y H, Liao X W, Chen E H. An implicit semantic enhanced fine-grained fake news detection method based on large language models[J]. Journal of Computer Research and Development, 2024, 61(5): 1250-1260.

[16] Chen Y, Li D, Zhang P, et al. Cross-modal ambiguity learning for multimodal fake news detection[C]//Proceedings of the ACM Web Conference 2022. New York: ACM Press, 2022: 2897-2905.

[17] Yang H Y, Zhang J J, Zhang L, et al. MRAN: Multimodal relationship-aware attention network for fake news detection[J]. Computer Standards & Interfaces, 2024, 89: 103822.

[18] Yang H Y, Zhu Y P, Luo J C, Nomaguchi H, Su C H, Susilo W. FusionVul: A multimodal feature fusion framework for source code vulnerability detection[J]. Journal of Systems and Software, 2026, 241: 112992.

[19] Xia J, Jia W, Li P, et al. Propagation Motifs as Codewords for Fake News Detection[J]. IEEE Transactions on Computational Social Systems, 2026, 13(1): 1015-1027.

[20] Cui C, Jia C. Propagation tree is not deep: Adaptive graph contrastive learning approach for rumor detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024: 73-81.

[21] Sharma S, Li Y. Fake news detection using temporal snapshots in

graph neural networks[C]//2025 International Conference on Computing, Networking and Communications. Piscataway: IEEE Press, 2025: 62-66.

[22] Wei L, Hu D, Zhou W, et al. Towards Propagation Uncertainty: Edge-Enhanced Bayesian Graph Convolutional Networks for Rumor Detection[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 3845-3854.

[23] Abduljaleel I Q, Ali I H. Deep learning and fusion mechanism-based multimodal fake news detection methodologies: A review[J]. Engineering, Technology & Applied Science Research, 2024, 14(4): 15665-15675.

[24] Khattar D, Goud J S, Gupta M, et al. MVAE: Multimodal variational autoencoder for fake news detection[C]//The World Wide Web Conference. New York: ACM Press, 2019: 2915-2921.

[25] Zhou X, Wu J, Zafarani R. SAFE: Similarity-aware multi-modal fake news detection[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer, 2020: 354-367.

[26] Wu G, Wang B, Li X, et al. Contrastive Learning Based on Feature Enhancement for Multi-modal Fake News Detection[C]//2024 43rd Chinese Control Conference. Piscataway: IEEE Press, 2024: 7610-7615.

[27] Zheng J, Zhang X, Guo S, et al. MFAN: Multi-modal feature-enhanced attention networks for rumor detection[C]// Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI). 2022: 2413-2419.

[28] Zhang Q, Liu J, Zhang F, et al. Natural Language-centered Inference Network for Multi-modal Fake News Detection[C]// Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI). 2024: 2542-2550.

[29] Xu F, Zeng L, Zou B, et al. CLFFRD: Curriculum learning and fine-grained fusion for multimodal rumor detection[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: ELRA and ICCL, 2024: 3314-3324.

[30] Liu Y, Liu Y, Li Z, et al. Modality interactive mixture-of-experts for fake news detection[C]//Proceedings of the ACM on Web Conference 2025. New York: ACM Press, 2025: 5139-5150.

[31] Wang B, Gao Y, Liu T, et al. Human-Like multimodal fake news detection via reflective summarization and large‐small model collaboration [J]. IEEE Transactions on Neural Networks and Learning Systems, 2026, 36(8): 1-15.

[32] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 10012-10022.

[33] Carlsson O, Gerken J E, Linander H, et al. Heal-swin: A vision transformer on the sphere[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2024: 6067-6077.

[34] Zubiaga A, Liakata M, Procter R, et al. Analysing how people orient to and spread rumours in social media by looking at conversational threads[J]. PLOS ONE, 2016, 11(3): e0150989.

[35] Jin Z, Cao J, Guo H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs[C]//Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM Press, 2017: 795-816.

[36] Ziari M, Charkari N M. Rumor detection and propagation on social networks: A survey[J]. Expert Systems with Applications, 2026, 295: 128798.

[37] Liu Y, Shen X, Zhang Y, et al. A systematic review of machine learning approaches for detecting deceptive activities on social media: methods, challenges, and biases[J]. International Journal of Data Science and Analytics, 2025, 20(7): 6157-6182.

谢丽霞（1974− ），女，中国民航大学教授，主要研究方向为网络信息安全、网络安全态势分析与评估、软件漏洞分析。

罗丹（2000− ），女，中国民航大学硕士生，主要研究方向为网络与信息安全。

杨宏宇（1969− ），男，博士，中国民航大学教授、博士生导师，主要研究方向为网络与系统安全、软件安全检测、网络安全态势感知。